Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth

In microbial ecology studies, the most commonly used ways of investigating alpha (within-sample) diversity are either to apply count-only measures such as Simpson's index to Operational Taxonomic Unit (OTU) groupings, or to use classical phylogenetic diversity (PD), which is not abundance-weighted. Although alpha diversity measures that use abundance information in a phylogenetic framework do exist, but are not widely used within the microbial ecology community. The performance of abundance-weighted phylogenetic diversity measures compared to classical discrete measures has not been explored, and the behavior of these measures under rarefaction (sub-sampling) is not yet clear. In this paper we compare the ability of various alpha diversity measures to distinguish between different community states in the human microbiome for three different data sets. We also present and compare a novel one-parameter family of alpha diversity measures, BWPD θ \operatorname{BWPD}_\theta , that interpolates between classical phylogenetic diversity (PD) and an abundance-weighted extension of PD. Additionally, we examine the sensitivity of these phylogenetic diversity measures to sampling, via computational experiments and by deriving a closed form solution for the expectation of phylogenetic quadratic entropy under re-sampling. In all three of the datasets considered, an abundance-weighted measure is the best differentiator between community states. OTU-based measures, on the other hand, are less effective in distinguishing community types. In addition, abundance-weighted phylogenetic diversity measures are less sensitive to differing sampling intensity than their unweighted counterparts. Based on these results we encourage the use of abundance-weighted phylogenetic diversity measures, especially for cases such as microbial ecology where species delimitation is difficult.

Connor O. McCoy and Frederick A. Matsen IV*
Fred Hutchinson Cancer Research Center
1100 Fairvew Ave. N
Seattle, WA 98109

*Corresponding author: matsen@fhcrc.org

1. INTRODUCTION

It is now well accepted that incorporating phylogenetic information into alpha (single-sample) and beta (between-sample) diversity measures can be useful in a variety of ecological contexts. Phylogenetic equivalents of all of major alpha diversity measures have been developed. Starting with Faith's original definition of phylogenetic diversity (Faith, 1992), which generalizes species count, there are now phylogenetic generalizations of the Simpson index to Rao's quadratic entropy (Rao, 1982; Warwick and Clarke, 1995), the Shannon index to phylogenetic entropy (Allen et al., 2009), and the Hill numbers to ${}^{q}\mathrm{D}(\mathrm{T})$ (Chao et al., 2010). Phylogenetic diversity itself has been extended to incorporate taxon counts (Barker, 2002) and proportional abundance (Vellend et al., 2011). There have also been abundance-weighted measures that explicitly measure phylogenetic community structure (Fine and Kembel, 2011), or an "effective number of species" (Chao et al., 2010). Many diversity measures can be tidily expressed in the framework of Leinster and Cobbold (2012), although the expression of phylogenetic diversity measures for non-ultrametric trees is complex.

In this paper we use three example human microbiome datasets to demonstrate the utility of abundance-weighted phylogenetic diversity measures. We also introduce a one-parameter family interpolating between classical PD and an abundance-weighted generalization. We call the parameter θ and denote the one-parameter family BWPD_θ; BWPD_0 is classical PD, whereas BWPD_1 is balance-weighted phylogenetic diversity, effectively $\mathrm{PD}_{\mathrm{aw}}$ of Vellend et al. (2011). Intermediate values of θ allow a partially-abundance-weighted compromise. Such a compromise has recently been shown to be useful for measuring beta diversity, with the introduction of a one-parameter family of "generalized UniFrac" measures (Chen et al., 2012). We use the name Balance Weighted Phylogenetic Diversity as described below because there are a variety of abundance weighted phylogenetic diversity measures. We compare the behavior of PD measures, including BWPD_θ, under various levels of sampling using theory and example data sets.

2. MATERIALS AND METHODS

2.1. **Datasets.** We apply the methods described below to three previously described 16S rRNA surveys of the human microbiome. The

first two datasets are composed of samples from "normal" and dysbiotic microbial communities, where previous studies have associated changes in diversity with changes in health. The third dataset investigates the changes of the skin microbiome through time.

2.1.1. *Bacterial vaginosis.* First, we reanalyze a pyrosequencing dataset describing bacterial communities from women being monitored in a sexually transmitted disease clinic for bacterial vaginosis (BV). BV has previously been shown to be associated with increased community diversity (Fredricks et al., 2005). For this study, swabs were taken from 242 women from the Public Health, Seattle and King County Sexually Transmitted Diseases Clinic between September 2006 and June 2010 of which 220 samples resulted in enough material to analyze (Srinivasan et al., 2012).

Selection of reference sequences and sequence preprocessing were performed using the methods described in (Srinivasan et al., 2012). 452,358 reads passed quality filtering, with a median of 1,779 reads per sample (range: 523–2,366).

2.1.2. *Oral periodontitis.* We also utilize sequence data from a study of subgingival communities in 29 subjects with periodontitis, along with an equal number of healthy controls (Griffen et al., 2011a). The publication analyzing this dataset showed increased community diversity in samples from diseased patients compared to healthy controls. Raw sequences were filtered, retaining only those reads with: a mean quality score of at least 25, no ambiguous bases, at least 150 base pairs in length, and an exact match to the sequencing primer and barcode. A total of 759,423 reads passed quality filtering, with a median of 8,320 reads per sample (range: 4,096–14,319).

As the phylogenetic placement method used below to calculate our measures requires a reference tree and alignment, we created a tree with FastTree 2.1.4 (Price et al., 2010) using the alignment and accompanying taxonomic annotation from the curated CORE database of oral microbiota (Griffen et al., 2011b).

2.1.3. *Skin microbiome through time.* Our third data set is a study of skin microbial diversity through adolescence Oh et al. (2012). Aligned sequences were obtained directly from the authors, although sequence data is available under the accession numbers [GQ000001] to [GQ116391] and can be accessed through BioProject ID 46333.

2.2. **Balance-weighted phylogenetic diversity.** In this section we introduce BWPD_θ, our one-parameter family interpolating between classical PD and fully balance-weighted phylogenetic diversity. We

will primarily consider so-called *unrooted* (Pardi and Goldman, 2007) phylogenetic diversity, which does not necessarily include the root. The case of *rooted* phylogenetic diversity can be calculated in a similar though simpler way as described below. Although we will primarily be working in an unrooted sense, it will be useful to use terminology that corresponds to the rooted case. For this reason, if the tree is not already rooted, assume an arbitrary root has been chosen; let the *proximal* side of a given edge be the side that contains the root and *distal* be the other.

We will describe BWPD_θ in terms of a phylogenetic tree T with leaves L, and a *contingency table* describing the number of observations of the organisms at the leaves in various samples. The contingency table has rows labeled with the leaves of T, and columns labeled by samples. In microbial ecology this is frequently known as an *OTU table*. The entry corresponding to a given leaf and a given sample is the number of times that leaf was observed in that sample.

The classical (unrooted) phylogenetic diversity of a given sample in this context is the total branch length of the tree subtended by the leaves in that sample.

The path to generalizing PD is to note that this can be expressed as a sum of branch lengths multiplied by a step function. Let $f(x)$ be the function that is one for $x > 0$ and zero otherwise. Let $g(x) = \min(f(x), f(1-x))$ and $D_s(i)$ be the fraction of reads in sample s that are in leaves on the distal side of edge i. Phylogenetic diversity can be then expressed as

$$\mathrm{PD}_\mathrm{u}(s) = \sum_i \ell_i \, g(D_s(i)) \tag{1}$$

That is, the sum of edge lengths in T which have reads from s on both the distal and proximal side.

Note that the step function g is the limit of a one-parameter family of functions (Fig. 1). Indeed, defining

$$g_\theta(x) = \min\left(x^\theta, (1-x)^\theta\right), \tag{2}$$

g is the pointwise limit on the closed unit interval of the g_θ as θ goes to zero. Thus our one-parameter generalization is

$$\mathrm{BWPD}_\theta(s) = \sum_i \ell_i \, g_\theta(D_s(i)). \tag{3}$$

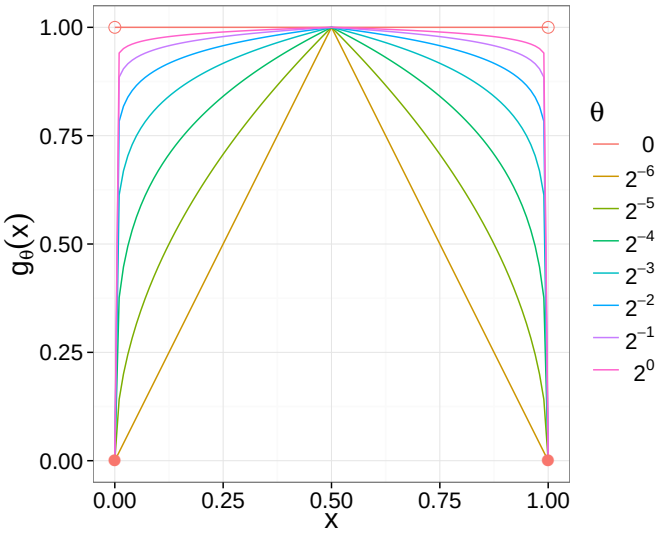

FIGURE 1. g_θ curves for various θ parameters. As θ goes to zero, the g_θ converge pointwise to g, which is 1 on the interior of the unit interval and 0 on the boundaries.

Note that when $\theta = 0$ this is PD and when $\theta = 1$ this is an abundance-weighted version of PD equivalent to executing the $\Delta\,\mathrm{nPD}$ recipe of Barker (2002) up to a multiplicative factor.

The rooted equivalent of (3) is

$$\mathrm{RBWPD}_\theta(s) = \sum_i \ell_i \left(D_s(i)\right)^\theta, \tag{4}$$

which interpolates between rooted PD and an abundance-weighted version. Vellend et al. (2011) describe a similar measure, $\mathrm{PD}_{\mathrm{aw}}$, which is equal to RBWPD_1 multiplied by the total number of branches in T.

We call BWPD_1 balance-weighted phylogenetic diversity because it weights edges according to the balance of read fractions on either side of an edge– edges with even amount of mass on either side are up-weighted, while edges with an uneven balance of mass are down-weighted. Indeed, if $|x-(1-x)|$ is taken to represent the imbalance of read fraction on either side of an edge, then $1-|x-(1-x)|$ can be taken to be a measure of balance; note that on the unit interval, $\min(x, 1-x) = 1-|x-(1-x)|$. Because a small x or an x close to 1 gives a small coefficient in the summation, small collections of

reads or small perturbations of the read distribution will not change the value of BWPD_1 appreciably.

2.3. **Calculation of PD measures in example applications.** Reads from the vaginal and oral studies were placed on a tree created from a curated set of taxonomically annotated reference sequences. As phylogenetic entropy and ${}^{q}\mathrm{D}(\mathrm{T})$ operate on a rooted phylogeny, reference trees were assigned a root taxonomically (Matsen and Gallagher, 2012). pplacer was run in posterior probability mode (using the `-p` and `--informative-prior` flags), which defines an informative prior for pendant branch lengths with a mean derived from the average distances from the edge in question to the leaves of the tree. The resulting set of placements were classified at the family rank using a hybrid classifier implemented in the guppy tool from the pplacer suite. The hybrid classifier assigns taxonomic annotations to sequences using the combination of a naïve Bayes classifier (Wang et al., 2007) with a phylogenetic classifier (Matsen et al., unpublished results). Any reads that could not be confidently classified to the family rank were not used in measures based on classification.

Full-length 16S sequences were available for the skin data, and so a more traditional tree-building approach was used. Representative OTUs were chosen for each site by clustering at 97% identity using USEARCH (Edgar, 2010), with trees built on OTU centroids using FastTree (Price et al., 2010). To conform with methods used in that paper, the naïve Bayes classifier (Wang et al., 2007) was used to infer genus-level classifications to taxonomically root the tree; in our case we used the RDP classifier v2.5. The contingency (OTU) tables generated by clustering were made available to our tools via the BIOM (McDonald et al., 2012) format.

PD_{u} (unrooted PD), phylogenetic quadratic entropy (Rao, 1982), phylogenetic entropy (Allen et al., 2009), and ${}^{q}\mathrm{D}(\mathrm{T})$ (Chao et al., 2010) were implemented for phylogenetic placements in the freely-available pplacer suite of tools (Matsen et al., 2010) (`http://matsen.fhcrc.org/pplacer`) in the subcommand guppy fpd. PD_{u} on rarefied phylogenetic placements was calculated using guppy rarefy.

Discrete measures of alpha diversity and richness were calculated on contingency tables obtained from clustering and taxonomic classification. Sequences were clustered into Operational Taxonomic Units (OTUs) at a 97% identity threshold using USEARCH 5.1 (Edgar, 2010). Similar results were observed when clustering at 95% identity (results not shown). OTU counts and family-level taxon counts were then rarefied to the read count of the specimen in the dataset with

the fewest sequences in R 2.15.1 (R Development Core Team, 2012) using the vegan package (Oksanen et al., 2012). We obtained values for the Simpson (1949) and Shannon (1948) diversity indices, as well as the Chao1 (Chao, 1984) and ACE (Chao and Lee, 1992) measures of species richness using vegan functions diversity and estimateR.

2.4. **Comparative analysis of alpha diversity measures.** To investigate the relation between various measures of alpha diversity, we calculated Pearson's r between all pairs of measures using the function rcorr from the R package Hmisc (Harrell Jr., 2012). We then performed hierarchical clustering with the R function hclust, using $d = 1 - r$ as the distance between two measures.

Association of each measure with clinical criteria for the first two data sets was evaluated by examining the accuracy of a logistic regression using the measure as the sole predictor of whether the sample came from a "normal" or dysbiotic subject. In the vaginal dataset, we assessed each measure's ability to predict whether a sample was from a subject positive for BV by Amsel's criteria, a clinical diagnostic method (Amsel et al., 1983). In the oral dataset, we assessed each measure's ability to predict whether a sample was from a healthy control, or a subject with periodontitis. Accuracy in predicting sample community state was assessed by leave-one-out cross-validation using the R package boot (Davison and Hinkley, 1997; Canty and Ripley, 2012).

For the vaginal dataset, we also calculated R^2 values using each measure individually as a predictor for sample Nugent score in a linear regression. The Nugent score provides a diagnostic score for BV which ranges from 0 (BV-negative) to 10 (BV-positive) based on presence and absence of bacterial morphotypes as viewed under a microscope (Nugent et al., 1991).

We calculated p-values to compare within- and between-stratification variability using R's built-in t.test function for the vaginal data, which had a binary stratification, and aov function for the oral and vaginal data sets. The vaginal dataset data was stratified by Amsel's criterion, the oral dataset by condition and sampling site, and the skin microbiome dataset by Tanner scale of physical development (Oh et al., 2012).

Plots were prepared with R base graphics and ggplot2 (Wickham, 2009).

2.5. **Evaluation of performance under rarefaction.** Phylogenetic placements were rarefied using the rarefy subcommand of the guppy tool

in the pplacer suite. Phylogenetic alpha diversity measures were calculated on the resulting rarefied placements as described above.

3. RESULTS

3.1. **Application to the human microbiome.**

3.1.1. *Vaginal microbiome.* Like Srinivasan et al. (2012) and many others in the field, we observe greater diversity in BV positive specimens using a variety of diversity and richness measures (Fig. S1). In particular, this is true for BWPD_θ for a variety of values of θ (Fig. S2).

In the vaginal data, phylogenetic measures of alpha diversity have better cross-validation accuracy for the Amsel classification and better correlation with the Nugent score than discrete OTU-based measures (Table 1). All measures were somewhat accurate in identifying community state, with even the worst performers classifying at least 70% of samples correctly. $\mathrm{BWPD}_{0.25}$, rarefied PD_u, PD_u, and phylogenetic entropy perform equally well predicting BV status. Correlation with Nugent score varies more widely, from 0.19 using Simpson (OTU) to 0.74 using $\mathrm{BWPD}_{0.25}$ or Simpson applied to family-level classifications. OTU-based measures rank in the bottom half of the measures tested, and below all phylogenetic measures. Phylogenetic diversity, which can be viewed as a measure of richness, outperforms discrete measures of richness, and most measures incorporating abundance.

In the hierarchical clustering of alpha measures on the vaginal data set, phylogenetic methods are separated from OTU-based methods (Fig. 2). BWPD_θ is similar to different extant phylogenetic alpha diversity measures for different θ. The Simpson and Shannon diversity measures cluster together, as do the ACE and Chao1 richness measures.

Fig. 3 shows values of BWPD_θ calculated before (x-axis) and after (y-axis) a single rarefaction to 523 sequences per sample. Samples for which the BWPD_θ value changes little lie close to the blue line, which shows the case of no difference between original and rarefied samples. Increasing θ, which corresponds to increased use of abundance information, reduces the change in BWPD_θ induced by rarefaction. Phylogenetic quadratic entropy and phylogenetic entropy both show behavior similar to BWPD_1, with rarefaction introducing little effect.

It might be possible to formalize a statement to this effect by computing the expectation of these alpha measures under rarefaction. However, computing the expectation for BWPD_θ under rarefaction

Measure	Nugent R^2	Amsel Accuracy	Amsel p-value
$\mathrm{BWPD}_{0.25}$	0.738	0.828	1.49E-35
Simpson (Family)	0.731	0.822	2.07E-33
Rarefied PD_{u}	0.731	0.828	6.81E-35
Shannon (Family)	0.721	0.821	8.85E-33
$\mathrm{BWPD}_{0.5}$	0.703	0.823	2.16E-33
PD_{u}	0.696	0.832	1.26E-32
Phylo. entropy	0.679	0.832	1.56E-31
${}^{0.25}\mathrm{D}(\mathrm{T})$	0.677	0.818	8.74E-30
${}^{0.5}\mathrm{D}(\mathrm{T})$	0.662	0.814	5.35E-29
Phylo. quad. entropy	0.647	0.811	7.70E-30
BWPD_{1}	0.610	0.796	5.66E-28
Chao1 (Family)	0.610	0.823	9.79E-24
Chao1 (OTU)	0.450	0.758	1.61E-19
ACE (OTU)	0.422	0.763	6.86E-20
Shannon (OTU)	0.380	0.754	6.73E-16
Simpson (OTU)	0.192	0.700	1.36E-07
ACE (Family)	0.088	0.666	1.51E-01

TABLE 1. Correlation and predictive performance of the various alpha diversity measures, ordered by decreasing R^2 value. Nugent R^2: R^2 value using the measure as a predictor, and the Nugent score as response in a linear model. Amsel accuracy: proportion of specimens with correct BV classification under a leave-one-out cross-validation. Amsel p-value: p-value from a two-sample t-test on values stratified by BV classification. "OTU" designates the measure applied to 97% clustering groups, and "Family" designates taxonomic classification at the family level. Measures described in main text.

does not appear to be straightforward: the methods of Dremin (1994) might be applicable in this setting, however, even the integer moments of the hypergeometric distribution are complicated and the non-integer moments are bound to be very complex. We have, however, shown in the Appendix that the expectation of phylogenetic quadratic entropy under rarefaction to k sequences assigned to the tips of a phylogenetic tree is

$$\mathbb{E}[\mathrm{PQE}_k] = \frac{k-1}{kn(n-1)} \sum_i \ell_i d_i (n - d_i)$$

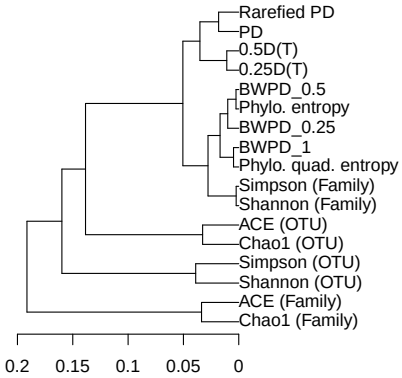

FIGURE 2. Dendrogram relating alpha diversity measures applied to the vaginal dataset.

where d_i is the number of sequences falling below edge i and ℓ_i is the length of edge i. This is almost identical to the unrarefied value of phylogenetic quadratic entropy, i.e.

$$\mathrm{PQE} = \frac{1}{n^2} \sum_i \ell_i d_i (n - d_i).$$

Thus it is not surprising to see that the expectation of PQE under rarefaction is very close to the original value (Fig. S3) for reasonably large k and n.

3.1.2. *Oral microbiome.* As previously observed by Griffen et al. (2011a), we find generally higher diversity in samples from diseased patients (Fig. 4). We evaluated the ability of each alpha diversity measure to predict whether a sample came from an individual with periodontitis, regardless of sample collection site, using the above methods.

In the oral dataset, phylogenetic alpha diversity measures incorporating abundance gave the best predictions of community state (Table S1, Fig. 4). In contrast, classical phylogenetic diversity was amongst the worst predictors; rarefaction did help, but rarefied PD still performed worse than phylogenetic measures taking abundance into account.

OTU-based methods and phylogenetic methods are not as separated in a hierarchical clustering as for the vaginal dataset (Fig. S6). However, many of the same pairings are present in both clusterings:

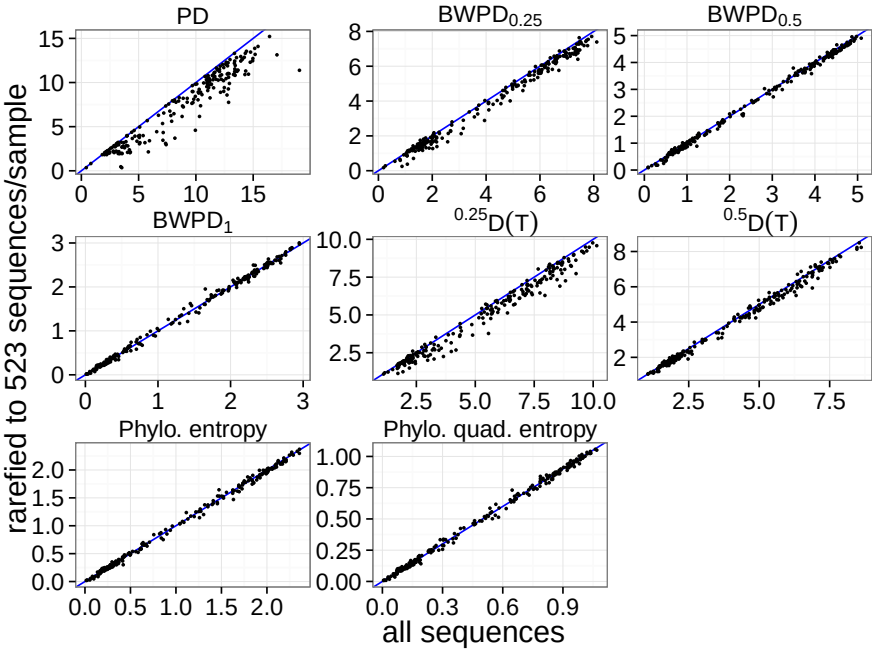

FIGURE 3. Comparison of rarefied and unrarefied values of various phylogenetic alpha diversity measures as applied to the vaginal dataset. The value of six alpha measures for each specimen using all available sequences is plotted on the x-axis. The value of the alpha measures for each specimen after a single rarefaction to 523 sequences (the smallest sequence count across specimens) is plotted on the y-axis. The $y = x$ line is shown in blue.

$\mathrm{BWPD}_{0.5}$ with PE, BWPD_{1} with QE, Simpson with Shannon, ACE with Chao1, and PD_{u} with rarefied PD. Interestingly, PD_{u}, rarefied PD_{u}, and $\mathrm{BWPD}_{0.25}$ all cluster with the discrete richness measures ACE and Chao1.

Like the vaginal dataset, incorporating abundance information decreases the effect of rarefaction on BWPD_{θ} values (Figs. S4, S5).

3.1.3. *Skin microbiome.* To further assess resolution and robustness of abundance weighted phylogenetic diversity measures, we considered skin microbiome data from a study by Oh et al. (2012). This

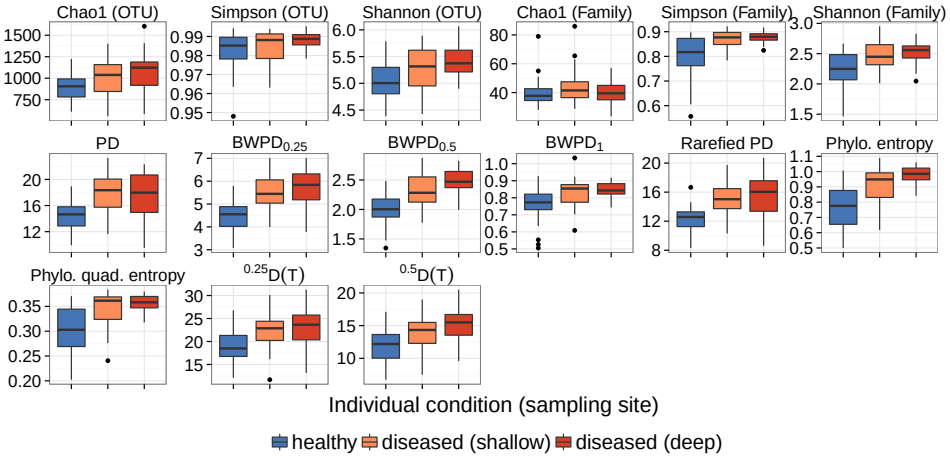

FIGURE 4. Comparison of diversity between samples from healthy controls, healthy sites of diseased patients, and diseased sites of diseased patients in the oral dataset, using different measures of alpha diversity. Top row: cluster-based methods. Bottom row: phylogenetic methods.

study tracked the changes of the skin microbiome through developmental stages. Because there are five Tanner stages, and they do not have a monotonic relationship with skin microbiome diversity (Oh et al., 2012), we focused on ANOVA p-values to see if the diversity measurements had small within-stage heterogeneity compared to between-stage heterogeneity. To compare the ANOVA p-values associated with the diversity measurements across the various data sets, we ranked the p-value of the diversity measures from lowest to highest for each data set individually. We averaged these ranks to gain an overall measure of performance. The results again show phylogenetic measures generally performing better than OTU-based measures (Tab. 5). In this case, a light weighting or no weighting of phylogenetic diversity by abundance performed better than full abundance-weighting. Note that we are not presenting these uncorrected p-values as evidence that there is an interesting relationship between skin microbiome and developmental stage, but rather are using p-values as a way of measuring within-stage heterogeneity compared to between-stage heterogeneity for the various measures.

	Ac	N	Pc	Vf	mean rank
$\mathrm{BWPD}_{0.25}$	2.90e-02	1.29e-03	4.94e-03	1.71e-04	3.50
PD_{u}	2.72e-02	5.48e-03	8.62e-03	5.06e-04	5.25
${}^{0.25}\mathrm{D(T)}$	2.95e-02	1.24e-03	1.53e-02	3.86e-04	5.50
${}^{0.5}\mathrm{D(T)}$	3.03e-02	4.95e-04	2.92e-02	3.20e-04	5.75
$\mathrm{BWPD}_{0.5}$	6.53e-02	7.34e-05	1.43e-02	1.42e-03	6.75
${}^{0}\mathrm{D(T)}$	3.08e-02	2.37e-03	9.77e-03	8.88e-04	7.00
Chao1 (OTU)	2.97e-02	2.68e-03	9.14e-03	9.97e-03	7.00
Shannon (OTU)	7.09e-02	8.48e-02	1.23e-01	2.70e-05	8.00
Phylo. entropy	1.17e-01	2.31e-05	8.37e-02	1.03e-02	8.25
Phylo. quad. entropy	2.52e-01	6.31e-06	4.77e-01	1.55e-01	9.00
Simpson (OTU)	1.17e-01	3.68e-01	8.75e-01	1.15e-04	9.50
BWPD_{1}	3.11e-01	2.99e-05	6.45e-01	5.33e-01	10.25

TABLE 2. ANOVA p-values for various phylogenetic diversity statistics applied to the skin microbiome data of Oh et al. (2012). Rows are ordered by increasing mean rank across sites. The same site abbreviations are used as in their paper: Af, antecubital fossa; N, nares; Pf, popliteal fossa; Vf, volar forearm.

3.1.4. *Applications summary.* In all three of the data sets investigated, abundance-weighted phylogenetic diversity measures showed good performance to distinguish between community states: between "normal" and dysbiotic samples in the oral and vaginal microbiomes, and between developmental stages in the skin microbiome. Notably, the best distinguishing measure in each dataset was both phylogenetic and abundance-weighted. BWPD_θ, our new family of abundance-weighted phylogenetic diversity measures, was highly correlated with clinical status although the value of θ most associated with community state varied. On the vaginal and oral data sets intermediate values of θ for BWPD_θ provide the best correlation with clinical status. These results correspond to analogous results for beta diversity, where an intermediate exponent for "generalized UniFrac" was the most powerful (Chen et al., 2012).

4. Discussion

Phylogenetic alpha diversity measures were more closely related to community state than were discrete measures based on OTU clustering for the data sets investigated here. This result is especially interesting given that the Simpson index, the Shannon index, or counting applied to OTU tables are very common ways of characterizing microbial diversity (Fierer et al., 2007; Grice et al., 2009; Hill et al., 2003; Dethlefsen and Relman, 2011). As also noted by Aagaard et al. (2012), we find that measurements of diversity using taxonomic classification can be useful in describing communities, and in fact perform much better than the same measurements of diversity applied to OTU counts; however, this approach requires a taxonomically well characterized environment. Our results can be viewed as an experimental confirmation of the notion that incorporating similarity between species is important to get sensible measures of diversity, which has been advocated by many, including most recently by Leinster and Cobbold (2012).

We find that classical phylogenetic diversity is sensitive to sampling depth, underestimating the true value in small samples. Biases have also been described for diversity measures using OTU tables (Gihring et al., 2012). In contrast, we observe that some abundance-weighted phylogenetic measures are relatively robust to varying levels of sampling.

As of the publication of this paper, no abundance-weighted phylogenetic alpha diversity measures are implemented in either mothur (Schloss et al., 2009) or QIIME (Caporaso et al., 2010), two of the most popular tools for analysis of microbial ecology data. Although the fact that abundance-weighted phylogenetic diversity measures performed best for the three data sets investigated here does not imply that they are best in general, we suggest that abundance-weighted phylogenetic measures be given greater consideration for microbial ecology studies. For this to happen, implementations in commonly used microbial ecology software packages will be needed, in addition to our implementation and that of the picante R package (Kembel et al., 2010).

5. Acknowledgements

The authors would like to thank Steven Kembel for encouragement and guidance, Steven N. Evans for probability consultation,

and David Nipperess for an interesting dialog concerning phylogenetic diversity and rarefaction. The Segre lab at the NIH, in particular Sean Conlan and Julia Oh, were very generous and helpful with the skin data. This work would not have been possible without an ongoing collaboration with David Fredricks, Noah Hoffman, Martin Morgan, and Sujatha Srinivasan at the Fred Hutchinson Cancer Research Center. This work was supported in part by NIH R01 HG005966-01.

REFERENCES

K. Aagaard, K. Riehle, J. Ma, N. Segata, T.A. Mistretta, C. Coarfa, S. Raza, S. Rosenbaum, I. Van den Veyver, A. Milosavljevic, D. Gevers, C. Huttenhower, J. Petrosino, and J. Versalovic. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLOS ONE*, 7(6):e36466, 2012.

B. Allen, M. Kon, and Y. Bar-Yam. A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *The American Naturalist*, 174(2):236–243, 2009.

R. Amsel, P.A. Totten, C.A. Spiegel, K. Chen, D. Eschenbach, and K.K. Holmes. Nonspecific vaginitis: diagnostic criteria and microbial and epidemiologic associations. *The American Journal of Medicine*, 74(1):14–22, 1983.

G.M. Barker. Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biological Journal of the Linnean Society*, 76(2):165–194, 2002.

A. Canty and B.D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2012. R package version 1.3-5.

J.G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Peña, J.K. Goodrich, J.I. Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010.

A. Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):pp. 265–270, 1984.

A. Chao and S.M. Lee. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, pages 210–217, 1992.

A. Chao, C.H. Chiu, and L. Jost. Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1558):3599–3609, 2010.

J. Chen, K. Bittinger, E.S. Charlson, C. Hoffmann, J. Lewis, G.D. Wu, R.G. Collman, F.D. Bushman, and H. Li. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16):2106–2113, 2012.

A.C. Davison and D.V. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997. ISBN 0-521-57391-2.

L. Dethlefsen and D.A. Relman. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4554, 2011.

I.M. Dremin. Fractional moments of distributions. *JETP letters c/c of Pis'ma v zhurnal eksperimental'noi teoreticheskoi fiziki*, 59:585–585, 1994.

R.C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

D.P. Faith. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10, 1992.

N. Fierer, M. Breitbart, J. Nulton, P. Salamon, C. Lozupone, R. Jones, M. Robeson, R.A. Edwards, B. Felts, S. Rayhawk, et al. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology*, 73(21):7059–7066, 2007.

P.V.A. Fine and S.W. Kembel. Phylogenetic community structure and phylogenetic turnover across space and edaphic gradients in western amazonian tree communities. *Ecography*, 34(4):552–565, 2011.

D.N. Fredricks, T.L. Fiedler, and J.M. Marrazzo. Molecular identification of bacteria associated with bacterial vaginosis. *New England Journal of Medicine*, 353(18):1899–1911, 2005.

T.M. Gihring, S.J. Green, and C.W. Schadt. Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environmental Microbiology*, 2012.

E.A. Grice, H.H. Kong, S. Conlan, C.B. Deming, J. Davis, A.C. Young, G.G. Bouffard, R.W. Blakesley, P.R. Murray, E.D. Green, et al. Topographical and temporal diversity of the human skin microbiome. *Science*, 324(5931):1190–1192, 2009.

A.L. Griffen, C.J. Beall, J.H. Campbell, N.D. Firestone, P.S. Kumar, Z.K. Yang, M. Podar, and E.J. Leys. Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *The ISME Journal*, 2011a.

A.L. Griffen, C.J. Beall, N.D. Firestone, E.L. Gross, J.M. DiFranco, J.H. Hardman, B. Vriesendorp, R.A. Faust, D.A. Janies, and E.J. Leys. CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. *PLOS ONE*, 6(4):e19051, 2011b.

F.E. Harrell Jr. *Hmisc: Harrell Miscellaneous*, 2012. URL `http://CRAN.R-project.org/package=Hmisc`. R package version 3.9-3.

T.C.J. Hill, K.A. Walsh, J.A. Harris, and B.F. Moffett. Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology*, 43(1):1–11, 2003.

S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26(11):1463–1464, 2010.

Tom Leinster and Christina A Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012.

F.A. Matsen and A. Gallagher. Reconciling taxonomy and phylogenetic inference: formalism and algorithms for describing discord and inferring taxonomic roots. *Algorithms for Molecular Biology*, 7 (1):8, 2012.

F.A. Matsen, R.B. Kodner, and E. Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1):538, 2010.

D. McDonald, J.C. Clemente, J. Kuczynski, J.R. Rideout, J. Stombaugh, D. Wendel, A. Wilke, S. Huse, J. Hufnagle, F. Meyer, et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Giga Science*, 1(1):1–6, 2012.

R.P. Nugent, M.A. Krohn, and SL Hillier. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *Journal of Clinical Microbiology*, 29(2):297–301, 1991.

Julia Oh, Sean Conlan, E Polley, Julia A Segre, Heidi H Kong, et al. Shifts in human skin and nares microbiota of healthy children and adults. *Genome medicine*, 4(10):1–11, 2012.

J. Oksanen, F.G. Blanchet, R. Kindt, P. Legendre, R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M.H.H. Stevens, and H. Wagner. *vegan: Community Ecology Package*, 2012. URL `http://CRAN.R-project.org/package=vegan`. R package version 2.0-4.

F. Pardi and N. Goldman. Resource-aware taxon selection for maximizing phylogenetic diversity. *Systematic Biology*, 56(3):431–444,

2007.

M.N. Price, P.S. Dehal, and A.P. Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3): e9490, 2010.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`.

C.R. Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982.

P.D. Schloss, S.L. Westcott, T. Ryabin, J.R. Hall, M. Hartmann, E.B. Hollister, R.A. Lesniewski, B.B. Oakley, D.H. Parks, C.J. Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75 (23):7537–7541, 2009.

C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(1):379–423, 1948.

E.H. Simpson. Measurement of diversity. *Nature*, 163(4148):688, 1949.

S. Srinivasan, N.G. Hoffman, M.T. Morgan, F.A. Matsen, T.L. Fiedler, R.W. Hall, F.J. Ross, C.O. McCoy, R. Bumgarner, J.M. Marrazzo, et al. Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLOS ONE*, 7(6):e37818, 2012.

M. Vellend, W.K. Cornwell, K. Magnuson-Ford, and A. Mooers. Measuring phylogenetic biodiversity. In *Biological Diversity: Frontiers in Measurement and Assessment*, A.E. Magurran and B.J. McGill, editors, pages 194–207. Oxford University Press, 2011.

Q. Wang, G.M. Garrity, J.M. Tiedje, and J.R. Cole. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16): 5261–5267, 2007.

R.M. Warwick and K.R. Clarke. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, 129(1):301–305, 1995.

H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.